Mining the News with Semantic Press

Eugenio Picchi, Eva Sassolini, Sebastiana Cucurullo, Francesca Bertagna
Istituto di Linguistica Computazionale (CNR-ILC)
Consiglio Nazionale delle Ricerche, Pisa, Italy
{eugenio.picchi|eva.sassolini|nella.cucurullo|francesca.bertagna}@ilc.cnr.it

Abstract

In this paper, we present Semantic Press, a tool for automatic press review based on text mining technologies and tailored to meet the requirements of eGovernment and eParticipation. The paper first provides a general description of the applicative exigencies that emerge from the eParticipation and eGovernment sectors. Then, an introduction of the general framework (the so called Linguistic Miner) for the automatic analysis and classification of textual content is provided. The core of the paper is the description of the tool for the analysis and presentation of newspapers content, its underlying technologies and final functionalities.
Index Terms: text mining, press review

1. Introduction

eParticipation is the extension and transformation of participation in political deliberation and decision-making processes through information and communication technologies (ICT). The notion is complementary the eGovernment one, which concerns more the use of ICT to improve and to innovate the quality of services offered by Public Administration to citizens.
Language plays a fundamental role in eParticipation, since it is the medium through which all the communication takes place: it is the language we find in institutional sites to explain to citizens how to obtain a particular service, it is the language of political discourse, the language of people expressing political opinions on a non official forum. NLP can be an instrument to deal with all these types of messages in an automatic or semiautomatic way.
A very sensitive issue for eParticipation and eGovernment is the necessity, for citizens but also for professionals of politics and consensus formation, to know salient facts and features, hidden in very large quantity of data, which standout for their frequency: this allows deriving interesting and constantly updated information about trends, tendencies and most important topics in a given period. For this kind of exigencies, ignited by the availability of huge quantity of information, constantly changing and dislocated on a great number of web sites, Text Mining techniques are very useful and promising.
In this paper we present Semantic Press, a tool for automatic press review based on text mining technologies and tailored to meet the requirements of eGovernment and eParticipation. The paper first provides a general description of the applicative exigencies that emerge from the eParticipation and eGovernment sectors. Then, an introduction of the general framework (the so called Linguistic Miner) for the automatic analysis and classification of textual content is provided. The core of the paper is the description of the tool for the analysis and presentation of newspapers content, its underlying technologies and final functionalities.

2. Linguistic Miner

Semantic Press is one of the applications derived by the so-called Linguistic Miner [1], a project started in 2003 with the aim of developing a framework for the automatic extraction of linguistic knowledge from very large amounts of texts (from different sources and in different formats) to be exploited in didactic, editorial and cultural products.
Building the Linguistic Miner involves two fundamental steps: first of all, the data are gathered, then they are linguistically analysed to be further processed and classified. The first step produced a repository (a “mine”) of around 200 millions words, together with an automatic topic classification of texts. This was achieved by exploiting procedures for the upgrade and augmentation of textual data in the “mine” and for the automatic acquisition from the Web through spider technology, both with periodic updating and also by means of user-defined paths. The Mine is thus constantly augmented in size.
The second step consists in the automatic linguistic processing of the textual material, by using modules of the PiSystem [2], an integrated framework for the treatment of textual and lexical material, where the most important module is the DBT (Data Base Testuale, Textual Data Base). The most effective procedures for further analysis of texts are POS tagging and lemmatization, which have been performed over 90% of the whole repository.
Many are the frameworks in which text mining techniques are applied and exploited (such as Inxight's LinguistX [3], IBM's Intelligent Miner [4], TextWise etc.). In this scenario, the Linguistic Miner stands out for its being based on tools and basic technologies developed to obtain good linguistic analysis as support of the entire application. Linguistic Miner is specifically tailored for analyzing Italian, but it is obviously open to other languages.
In the last year, Linguistic Miner has been addressed to meet the exigencies of political and institutional bodies, such as Regione Toscana, which have expressed their interest to use and exploit a tool for the intelligent access to the flow of news and information provided by Italian newspapers available on the Internet. This aim is in line with the current interest of CNR-ILC to the themes of eParticipation and eGovernment, testified also by the participation to the DEMO-net project (http://www.demo-net.org/demo).

3. Semantic Press

Semantic Press specializes some of the functionalities of Linguistic Miner towards the analysis of information available in Italian on-line newspapers. Semantic Press can be reached at ULR http://serverdbt.ilc.cnr.it/edicola/. Semantic Press is different from other tools for automatic press review (such as Press Today, see http://test.presstoday.com/): as a matter of fact, it is not only a way to present and to incrementally store news and articles pertinent to different

sectors, but also a powerful tool, based on NLP and text mining techniques, for highlight emerging subjects, issues and words. Fig. 1 provides an overview of the entire system.
Every morning, Semantic Press downloads all the articles of the most important and most read Italian newspapers: Sole 24h, La Stampa, La Repubblica, Il Corriere della Sera, Il Messaggero, La Nazione, Reuters and La Voce. The acquisition procedure downloads not only the text and the title of the article, but also visits and saves all the textual material available in the web pages linked to the article. Some filters are activated, in order to avoid downloading of dossiers, tables and other not interesting sources. During the day, Semantic Press performs the updating of the articles, adding new information if available and handling cases of similarity between different versions of the same article. In this way, about 1200 new articles are stored every day.
The acquired textual material is saved and converted in an internal format based on the DBT specifications. When the article is saved, it is also classified according to a pre-defined set of ten topics (politics, finance, sport etc.). The classification is based on the classifying tags already present in the sources, which are normalized and mapped onto a shared and common classification scheme.

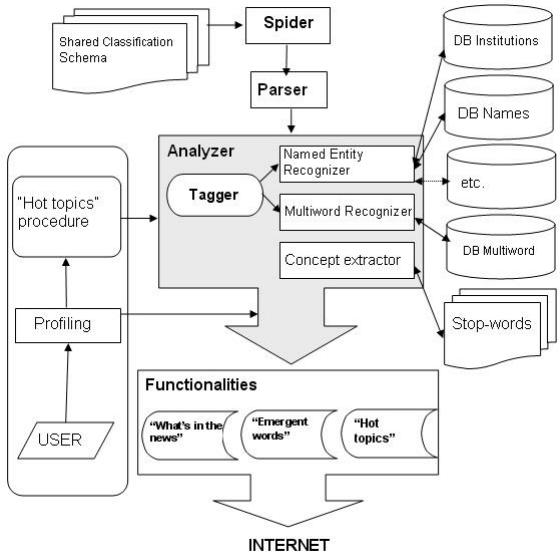

Fig. 1: "Semantic Press" system overview

Three basic technologies are used in the analysis phase: named entity recognition, multiword recognition and concept extraction. Different experiments have been carried out to evaluate the impact and performance of the basic technologies exploited, as well as an evaluation of the various strategies adopted.

The basic technologies are exploited by the application modules which provide results interesting for the users: the "What's in the news.." and the "Emergent Words.." functionalities. A particular step is represented by the so-called "Hot Topics" functionality, which performs classification on the basis of user-specific and user-defined exigencies.

4. System basic technologies

4.1. Named Entities Recognition

An important module of the system performs Named Entity Recognition on the bulk of knowledge acquired every day, extracting names of person, locations, addresses and organizations. These classes, which belong to the larger set of entities recognized within the context of international evaluation campaigns, such as MUC and EVALITA, are those more in line with the aim of a press review.
The approach is hybrid and exploits both external ad-hoc resources and consolidated techniques based on Support Vector Machine. The NERec module consists of a set of finite-state automata which benefit of ad-hoc databases containing more than 200.000 pre-classified named entities derived by on-line repositories (such as gazetteers, lists of names and surnames, lists of locations etc.). These databases are also the base of the dynamic recognition of new entities, resulting from the combination of only partially known entities and specific heuristics which provide a feedback on the original databases and constantly enrich them. In this way, the system is able to recognize the unknown Italian surname Andreoli by exploiting the co-occurrence with a known name, Alberto. Then, after the recognition, the "new" surname is added to the database. The NERec module also performs term disambiguation in case of ambiguous classes (typical the case geo-political entities) by exploiting ad-hoc statistical strategies. An important functionality is represented by the possibility to use lists of synonyms and variants, able to recognize, as a same entity, different forms: this is done, in particular, for the names of organizations, which are often alternatively present in form of acronym or whole name.

4.2. Multiword Recognition

This module is based on the analysis of very large corpora, both belonging to the Linguistic Miner and the repository of press articles; multiwords are extracted by exploiting pattern-matching techniques (the typical N-Adj and N-preposition-N patterns of Italian constructions) and filtered on the basis of the frequency distributions on the various sectors with respect to the reference corpora of the Linguistic Miner.

4.3. Concept Extraction

This basic functionality extracts all the terms which are above a given frequency threshold and are recognized as "semantically salient" terms. The module uses list of stop-words and other heuristics to determine the relevance of the term. In the specific case of the Hot Topics functionality (see section 5.3), the term extraction is carried out by means of the following process:

1. starting from a small set of pivot terms, a vocabulary is obtained based on mutual information;
2. the terms of the vocabulary are used to "weight" each single article;
3. the vocabulary is then enriched by exploiting the ranking of the archive of the news. In this way, the weight of each term and of each article is compared with the weight they have in the reference generic corpus.

Then, the articles are linguistically analysed in order to obtain texts annotated at lemma and PoS level.

5. System Functionalities

The textual analysis performed on the daily press allows the system to provide users with three functionalities: "Emergent Words", "What's in the News" and "Hot Topics". They are conceived as alternative views and access to the same content. Fig. 2 is a screenshot of the Semantic Press home page: the frame menu presents the set of available functionalities, while in the principal frame the various journalistic sources, the different themes and the relevant news are presented.

5.1. What's in the News

Semantic Press presents to the user the most important themes emerged from the automatic analysis of the daily news. User can know, each day, which are the most discussed themes, in form of name of persons (often politicians, but also people of the show business, important scientific or artistic personalities, protagonists of crime news etc.), events and facts, locations where important things are happening etc. Examples of possible topics, as they emerged from the Italian news in the current period, are *immigrants*, *tesoretto* (a revenue which exceeds the expectation), *welfare*, *Padoa Schioppa* (the Italian Minister of finance) etc.

Themes are selected, on the basis of statistical evidences, among the concepts highlighted during the textual analysis phase.

5.2. Emergent Words

Emergent words are automatically obtained by exploiting information on relative frequency of a term and its belonging to specific sectors. The aim is making emerge all the different words in a newspaper article, by providing a general overview of the vocabulary predominant in a particular day.

5.3. Hot Topics

A particular form of content classification is the one implemented in the functionality called Hot Topics, which allows the user to retrieve all the articles and news concerning a specific argument. First, by exploiting techniques based on mutual information and words co-occurrence, specific dictionaries are extracted by the bulk of information starting from a selection of "pivot terms". This allows, for example, deriving a dictionary of terms concerning sport starting from pivot terms like "Tour de France" and "cycling". Then, these argument-specific dictionaries are projected on the news repository and the system provides a ranking of the relevancy of each article to the given argument. The selection of the arguments is driven by user-specific interests: if a user wants to investigate a particular aspect present in the news in a given period, it can ask the service to report its evolution and behavior. The functionality is provided of a module for user profiling, which can direct the search and the extraction to specific interests of the user.

5.4. Mining the local news

Users are often interested in information of very local nature (see paragraph 7 for a generic use-case). For this reason, a specific functionality performs text mining on local news, offering a customized service to citizens living in different Italian cities.

This kind of "transversal" classification requires a specific treatment of the textual materials: first of all, the frequency thresholds established to select salient terms in the national news have to be modified. Moreover, it is not possible anymore to exploit the domains that classify the content in the national press: as a matter of fact, the on-line version of newspapers does not contain a reliable classification of the local news. Thus, the only classification provided in this case is the one supported by the "Hot Topics" functionality.

5.5. Web Alert

A specific functionality of the system is the one that sends an e-mail to the user each time news of his/her interest is published on the press. This functionality allows tracking the evolution of specific information in the news.

A specific feature of our Web Alert, with respect to similar available web solutions, is that it works by exploiting all the mining solutions implemented in the system, in particular the innovative classification strategies created as support of the "Hot Topics" functionality.

This means that the information will be found and announced not only in case of a simple keyword-based matching but also if the article is selected by projecting, on the incoming news, the dynamic, ad-hoc vocabulary extracted by the archive.

For example, if the user is interested on the situation in the Middle East, he/she will be "alerted" not only in case of an incoming article containing the string "Medio Oriente" (Middle East), but also if the news contain the words "Afghanistan" and "guerra" (war) or "Abu Mazen" and "Palestina".

6. Accessing normative texts

A further access modality is foreseen for normative text. In Semantic Press the aim is offering the wider range of access modalities and of prompt announcement of news. We want to obtain similar results also in the Legal domain, which is represented by a very particular type of text (the normative text). Specific modules for multiword recognition and concept extraction, more tailored to the analysis of this particular type of text, are used to access and browse laws and regulations. This new application is called Edicola Juris and works on an archive of legal texts derived by the "Gazzetta Ufficiale" (the official, periodic publication that collects all the new national Italian laws). In Edicola Juris, the extracted terminology (composed by single terms and multiwords) is used to help the user to restrict the scope of his/her search. For example (see Fig. 3) the user may ask all the articles of law concerning the "corte dei conti" (the Italian state audit court): the articles will be returned, together with the terminology calculated on the articles themselves. Each single term of the terminology may be used, as a sort of *facet*, to restrict the search and to obtain more precise results, for example articles that concern the "corte dei conti" and "contrasto all'evasione" (fight against tax evasion).

7. Use Case

"Profiling" is one of the most interesting features in Semantic Press: it allows the identification of specific needs of very different users.

Every levels of access allow the specialization of the informative offer, not only by means of the selection of the most suitable sources of information and of the sectors closer to the user exigencies but also by exploiting the classificatory capabilities of the "Hot Topics" functionality, which offers in this sense a high flexibility for user-specific needs.

An exemplifying user-scenario is the one of a service company that can offer to its users (often not "normal" citizens but rather other companies, typically public utilities companies) the possibility of obtaining very precise information on very particular sectors of interest. Public utilities companies rarely have the appropriate size and the personnel to be constantly "on the news": this is why they often ask an external service to provide them with the information of their interest. Often, the interesting information for them is of very "local" nature and very specific relevance. This type of information can hardly be restricted to a sector defined in aprioristic way. If we look at a public utilities company for refuse collection, for example, we see that salient information usually concern calls for bids and notice of modification in regulations of the specific sector. How can such specificity be successfully handled by an automatic system? Which are the words that can help us to circumscribe the sector we are trying to deal with? In a situation like this, the technology underlying the "Hot Topics" functionality may be of great help.

As a matter of fact, by using the "Hot Topics" option, the search is not restricted to single pivot words typical of the sector, but it is projected on an entire customized vocabulary, automatically created on the basis of the written material pertaining to the specific sector. The more specific is the sector, the more detailed and particular will be the vocabulary. In this way, users can find the information they are looking for, with a highly customized service.

8. References

[1] Picchi E., Ceccotti, M. G., Cucurullo, S., Sassi, M., Sassolini, E. (2004). Linguistic Miner: an Italian Linguistic Knowledge system. In Proceedings of LREC 2004

[2] Picchi E., Statistical Tools for Corpus Analysis: A Tagger and Lemmatizer for Italian, in Willy Martin, Willem Meijs, Margreet Elsemiek ten Pas, Piet van Sterkenburg & Piek Vossen (Eds.), Proceedings of Euralex '94, Amsterdam, The Netherlands, 1994.

[3] Effective Information Discovery, Supporting the Analytical Mission through Entity-Based, Semantic Discovery. An Inxight Federal Systems Group White Paper, November 2006.

[4] Installing the Intelligent Miner products: Modeling, Scoring, Visualization V8.2. IBM Manuals for DB2 Intelligent Miner Modeling V8.2.

.

Fig. 2: Semantic Press home page

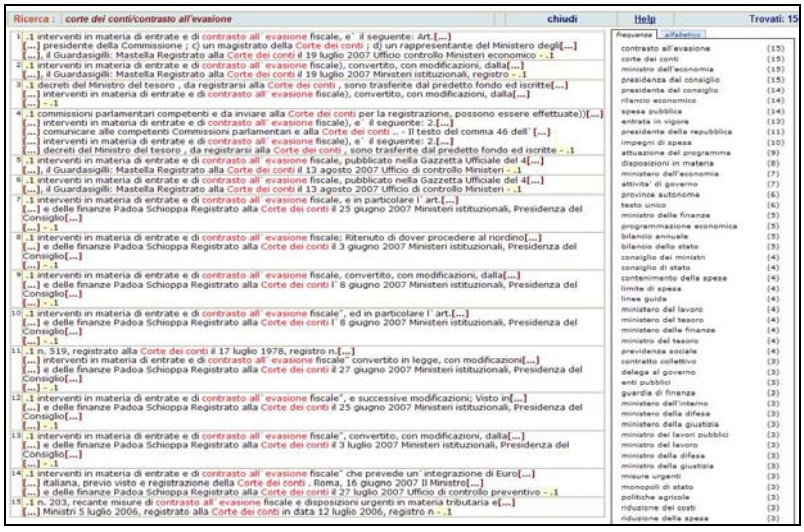

Fig. 3: Edicola Juris: analyzing normative text